Tilts and Bends Are To Be Cleared Away

		Page
7A.	A simple tilt assessor	7-2
7B.	That midmean flavor	7-8
7C.	Midmeans in tilt assessment	7-15
7D.	Introduction to bends and unbending	7-19
7E.	Assessing bend free of tilt	7-25
7F.	A simple bend assessor	7-27
7G.	Lessons about fitting in general	7-32
7H.	How far have we come?	7-35
	*****	
7K.	How to add	7-38
	*****	
7P.	IEDA processor features	7-40
	*****	
7R.	A few words about measurement	7-41
7S.	When computing is no pain	7-44
	*****	
7Y.	Problems	7-49

Location of tables and plots for 7

Number	Page
1	7-3
2	7-4
3	7-7
4	7-9
5	7-11
6	7-12
7	7-16
8	7-17
9	7-18
10	7-20
11	7-22
12	7-23
13	7-29
14	7-31
15	7-46
16	7-48

7

Tilts and Bends Are to be Cleared Away

Clearing away big -- or obvious -- effects so as to see smaller -- or camouflaged -- ones is a major task of exploratory data analysis. Indeed one can argue that almost every one of its techniques does this in one way or another. Residuals have had the fit subtracted so we can see what remains. When the typical values of two groups are compared, one with another, we can see small differences even when all typical values are very large. And so on.

The last step we took as part of a general program of improving plots was to flatten them. Operating graphically, we could do this quite well by eye. Operating numerically, as we also wish to do (not because it is harder but because complex cases tend to escape graphical treatment), we need to have some simple and effective techniques for choosing fits. Their description and use comes in this chapter.

When we operate graphically, we can try to straighten the picture out before flattening it. Numerical approaches are almost never as flexible as graphical ones. Accordingly we will want to have a procedure that flattens the data out whether or not it is straight -- whether or not it has been straightened. Here "flattening" must refer to the general run of the data, and not to its detailed behavior. Something that looks like a plausible set of residuals from a straight line is "flattened", though it may not be flat. Plot 30 of chapter 2 is an example.

Plot **1** shows a rather extreme example of curvature and plot **2** the results of "flattening" plot 1. Clearly plot 2 is far from being flat, so "flattening" is a bad term. Equally, it is poor use of words to talk about "the slope" of plot 1, since there is a very small slope toward the left of the plot and a very large slope to the right. Tilt is a short word, and conveys an appropriate feeling. We shall say that plot 2 is untilted, and that the slope of the line whose subtraction converts plot 1 into plot 2 -- thus clearing away this part of what is going on so that we can see better what remains -- is the tilt of plot 1.

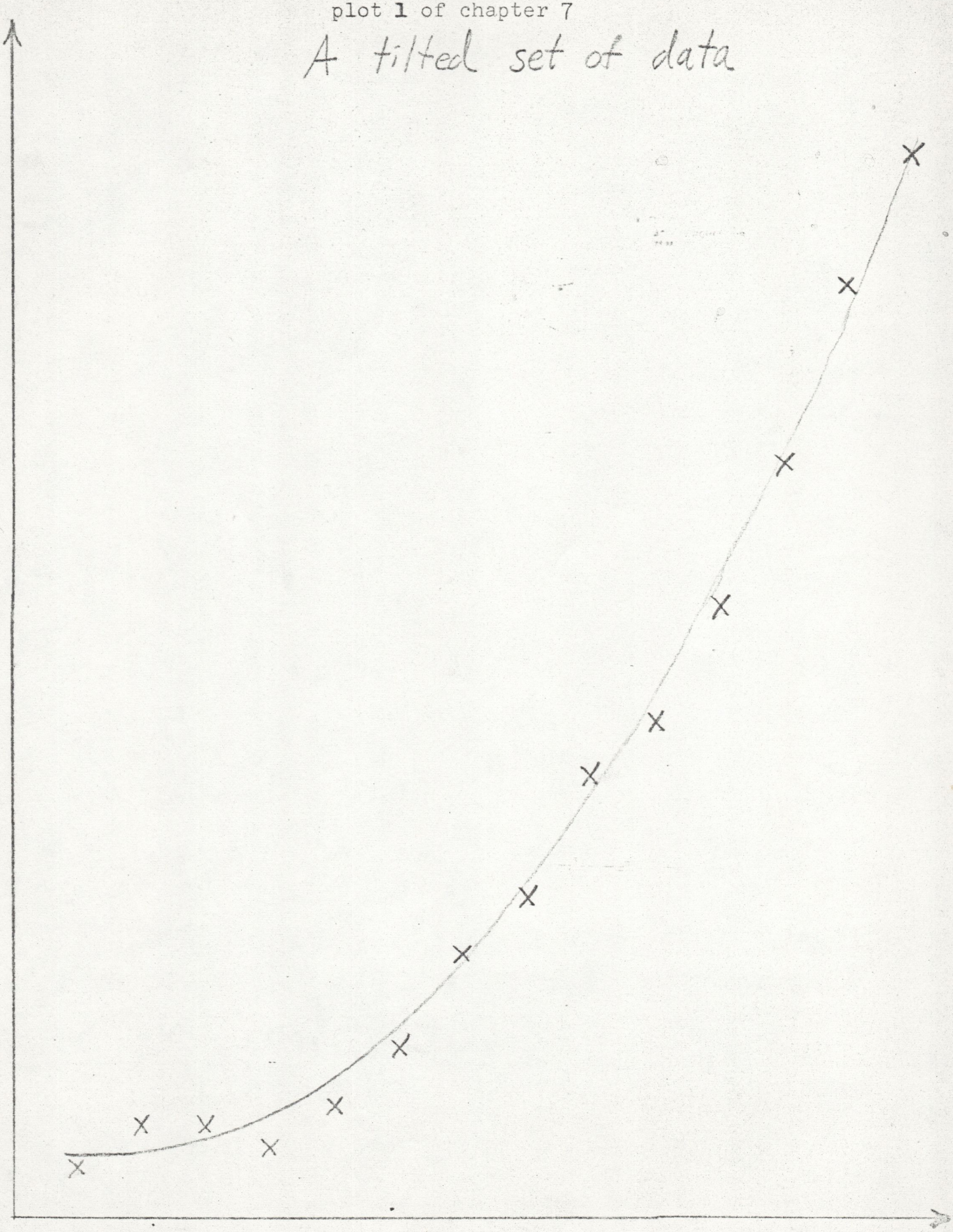
7A. <u>A simple tilt assessor</u>

A simple question that we often need to ask is: "How much is y tilted as a function of x?". This question we will take to mean: "What value of b leaves **y - bx** apparently untilted?". We shall take b as our assessment of tilt, and the values of **y - bx** as what we have been able to do about freeing y of tilt, so far as x is concerned. There can be a variety of different -- often only slightly different, usually not very much different -- definitions of "apparently untilted". For each such definition we shall find a comparably different value of b and thus a comparably different assessed value for the tilt and a comparably different set of apparently untilted values. This usually bothers us not at all.

In this whole introduction, our concern with tilt is like the woodsman's concern with bushes and fallen trees on a trail he plans to use:

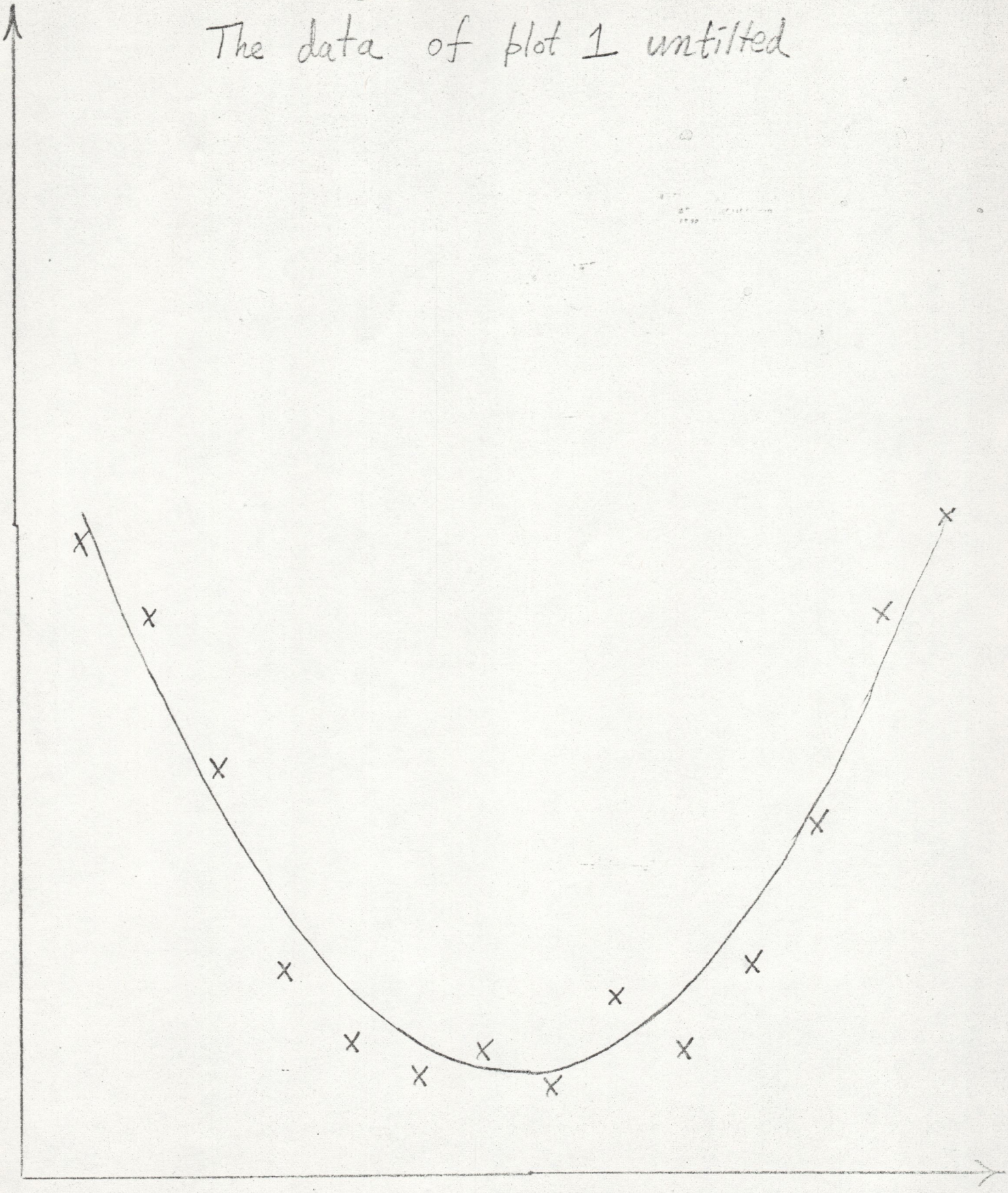
plot 1 of chapter 7

A tilted set of data



plot 2 of chapter 7

The data of plot 1 untilted



- We may be concerned to know there appears to be some tilt (though we may have known that there would be a tilt in this particular direction long before we collected the data).

- We may even want to know how much tilt there seems to be.

- We are almost certain, however, to want to clear the tilt out of our way.

We have already, in specific instances, cleared our way by using an eye-judged tilt to flatten our graphs. We frequently need to do something similar in fixing numbers for future analysis. Especially because we are going to use any tilt we assess in such a way, we will meet our needs quite well enough if we do a reasonably good job of assessing our tilt, without worry as to in what way, if any, our assessment might be best possible.

A simple and effective way to assess a tilt, especially when our x values are roughly equispaced, is to divide the observations into three portions -- low, middle, and high -- according to the values of x and then to form the means of x and y for the low portion, the high portion, and for all portions together. We shall write x_L, y_L and x_H, y_H and $\bar{x}, \bar{y}$ for these means. Then one choice is to take

$$\text{slope} = b = \frac{y_H - y_L}{x_H - x_L}$$

-- which is the slope of the line joining (x_L, y_L) to (x_H, y_H) -- as our slope, and agree that the fitted line should go through the overall mean $(\bar{x}, \bar{y})$, thus giving its equation the form

$$y = \bar{y} + b(x - \bar{x})$$

To make this approach operational we have only to say how to pick the three portions. A simple rule of thumb does quite well. To state it easily, let min x be the lowest value of x among our observations and let max x be the highest. Observe that one-third of the way from each toward the other falls at

$$\frac{2(\text{min } x) + (\text{max } x)}{3} \quad \text{and at} \quad \frac{(\text{min } x) + 2(\text{max } x)}{3}$$

The rule can then be put this way:

- include at least one-third of the interval from min x to max x in each end portion
- include at least one-quarter of the observations in each end portion (if this does not violate the first condition, otherwise include as close to one-quarter as you can)
- otherwise make each portion as small as you can.

Table 3 shows an example of the application of this procedure. (To keep the arithmetic under control, see "How to add" in Section 7K.) We could hardly have done a high quality job of fitting a line to this data <u>by arithmetic</u> with notably less labor.

table 3 of chapter 7

Fitting a Line by the Simple Rules

a) DATA AND FIT

(1) Years (= x) divided into three portions.

(2) Value of U. S. Imports of Merchandise (= y), in millions of dollars.

(3) and (4) Means of end portions.

(5) and (6) Overall means.

(1)	(2)	(3)	(4)	(5)	(6)
1918	3,031				
1919	3,904	1919.5	3,680		
1920	5,278				
1921	2,509				
1922	3,113				
1923	3,792			(1923.5)	(3,850.8)
1924	3,609				
1925	4,227				
1926	4,431				
1927	4,185	1927.5	4,186		
1928	4,091				
1929	4,039				

(7) Computation of slope: $b = \frac{4,186-3,680}{1927.5-1919.5} = 63.2$

(8) Fitted line = 3,851 + 63.2(x-1923.5).

b) CALCULATION OF RESIDUALS

Year	Imports	Fit	Residuals
1918	3031	3503.4	-472
19	3904	3566.6	+337
1920	5278	3629.8	+1648
1	2509	3693.0	-1184
2	3113	3756.2	-643
3	3792	3819.4	-27
4	3609	3882.6	-274
1925	4227	3945.8	+281
6	4431	4009.0	+422
7	4185	4072.2	+113
8	4091	4135.4	-44
9	4039	4198.6	-100

Example: 3503.4 = 3851 + 63.2(-5.5)

The corresponding residuals are plotted in plot 4, where we see -- if we had not noted in table 3 -- that there was an unusual perturbation near 1920. Clearly imports in 1920 were strikingly high. While the low value in 1921 is not so striking, it does appear possibly out of line.

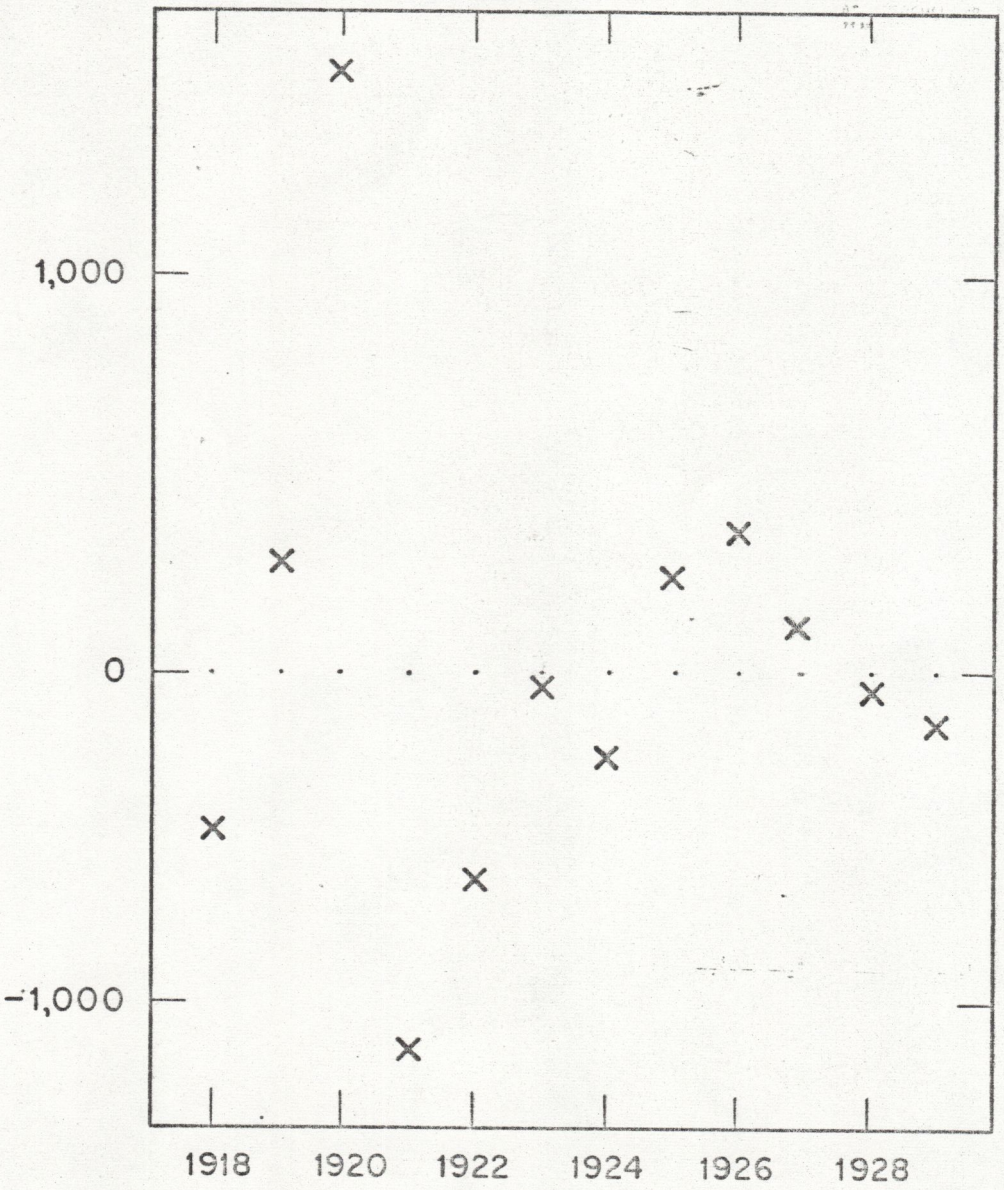
Should we do anything about this? The answer will depend greatly upon our interpretation of the situation -- and on our detailed purposes in fitting the line. If we regard 1920 as different from the more simply behaved years we are trying to describe by our fit -- or if we want to be sure that unusual years do not distort the residuals obtained for other years, we will wish to find some way to proceed that uses -- or is reminiscent of the use of -- midmeans. In the opposite circumstances we are likely to be content where we are.

7B. <u>That midmean flavor</u>

We could introduce midmeans by merely taking the y-values in the lower and higher portions according to x, forming their midmeans instead of their means, and using these values to fix the tilt. This might not be too bad, but is certainly not always good enough.

So long as we had only a batch of single values, taking a midmean was a simple operation. We could line up our values in order and drop off the end quarters without difficulty -- there was no serious choice of what order to use. Once each piece of data involves two numbers, one y and one x, things can become more complicated. We can order our data units -- conveniently called "points" or "data pairs" -- in many ways.

plot 4 of chapter 7



RESIDUALS FROM LINE FITTED IN TABLE 3

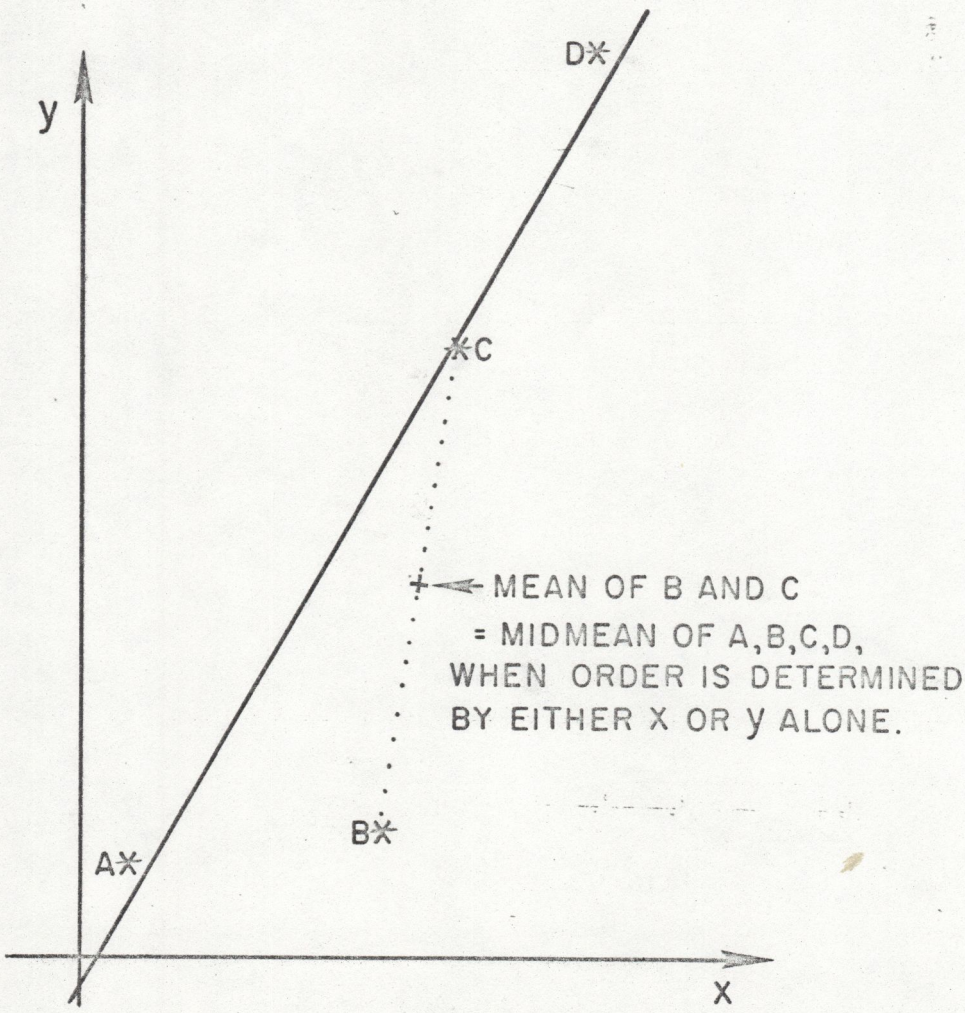
Ordering in accord with the x-values is surely possible. The y-values can be used, as could the values of $\sqrt{2}\cdot x + \pi\cdot y$ -- or the values of any other linear combination of x and y. Which we choose will have to depend upon the purpose for which the result is to be used. Indeed, knowing the broad character of the purpose may not be enough, its minor details may be very important.

At the moment, we are interested in midmeans of two portions of the data for a specific purpose -- finding a slope to be used for a fitted line -- a fitted line that will generate residuals for further examination. Our choice of an ordering should then reflect -- at least roughly -- the relative tendency of different points (= different data pairs) to tend to change this slope by increasing or decreasing it. Plot 5 shows quite clearly that the order given by x-values alone -- or by y-values alone -- need not provide a midmean with the properties we desire. In the situation shown, points A, C and D lie close to a line, while B is discrepant. Yet both the x-order and the y-order are $A < B < C < D$, so that the corresponding midmean -- the mean of B and C -- is far from the line. Using either x-selected midmeans or y-selected midmeans is not <u>always</u> satisfactory. We <u>may</u> have to do better.

Plot 6 shows, in the top panel, four points pulling on a contemplated line. The direction and, at least roughly, the relative strength of the pulls is clearly as indicated -- so long as we do not bring in the midmean flavor. The lower

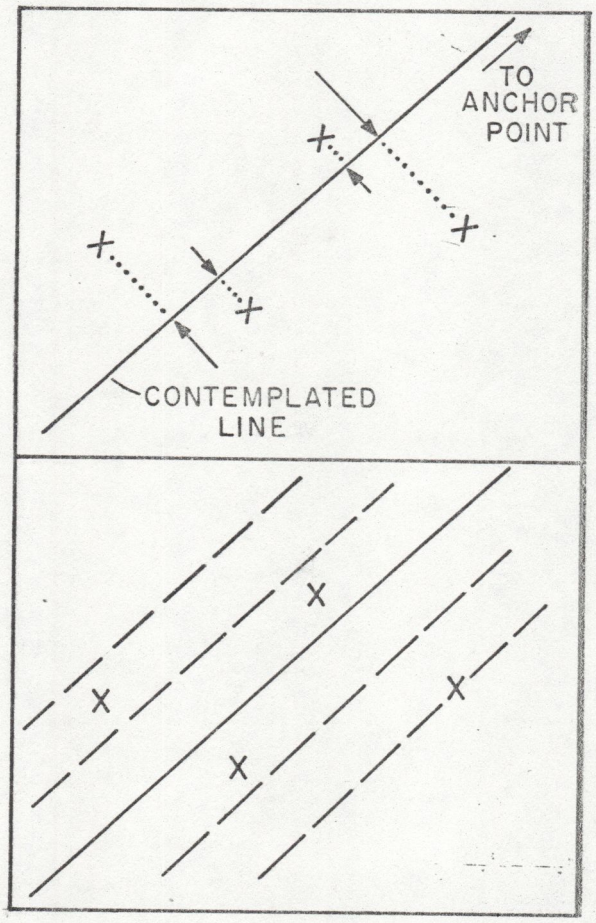
plot 5 of chapter 7

THE MISLEADING POSSIBILITIES OF X-SELECTED (OR y- SELECTED) MIDMEANS ILLUSTRATED.



plot 6 of chapter 7

FOUR POINTS PULLING ON A CONTEMPLATED LINE AND THE ORDERING INDUCED BY THE VALUES OF Y-BX



DASHED LINES CORRESPOND TO CONSTANT VALUES OF Y-BX, WHERE B IS THE SLOPE OF THE CONTEMPLATED LINE

panel shows that the ordering corresponding to a family of straight lines parallel to the contemplated line does a reasonably good job of reflecting the relative pulls of the various points.

The position is thus clear:

- if we know which direction the line will take, we know how to order the points in each portion -- thus if we know the slope, we can find the corresponding midmeans
- if we know the midmeans, we can find the slope.

This sort of chicken-and-egg situation is common -- not only in data analysis. Such an equation as

$$x^{17} = 5x^7 + 10{,}000$$

is a more extreme example. In this equation:

- if we know x^7 we can calculate $5x^7 + 10{,}000$ and find x^{17}.
- if we know x^{17}, we can find its 17th root, which will be x, and thence find x^7.

The circularity is quite similar.

One broad class of approaches to such problems is the use of repeated calculation, in which we make:

- a first calculation to obtain approximate answers;
- a second calculation, using these first approximations, to obtain second approximations;
- a third calculation, using the second approximations;
- and so on, as needed.

In such repeated - often called iterative - calculations, the first step is often (wisely) of one kind, while all the others are of another.

How should we apply these ideas to finding and using the midmeans of the low and high portions of our data? To answer this question we need to be concerned, separately, with the iterative step and the starting process.

If B is a contemplated (or "trial" or "previous approximation") value of the tilt, the natural iterative step is to calculate the "rough residuals" y - Bx, to order the data pairs by the values of these rough residuals, use this ordering to determine which data pairs are to be used to form the midmeans, find the midmeans and proceed. As we shall see in the next section, this process proceeds easily and smoothly.

How shall we start? If the iterative calculation is effective enough, the only important consequence of how we start is how many iterations we will need to make. (We always hope that one iteration after the start will suffice; we do not always have so pleasant a situation.) It is always easy to start with the B obtained by using the MEANS of the portions. This is usually not too bad a start, but if there are obvious "wild points" in one or both positions, it may be wise to begin by using means, but not means of the whole portions. Instead, we can use the means of what is left of each portion after any highly suspicious point (here, point = data pair) or points have been set aside. (We do, of course, bring these points back

into later steps, depending upon the midmean calculation to hold down their influence appropriately.)

7C. Midmeans in tilt assessment

The portion means that were used in the example worked out in Section 7A seemed not to be notably unreasonable. (The low value in 1921 tends to balance the very high value in 1920.) Accordingly we may as well use the slope found there as the starting slope for a calculation of residuals, and, thence, midmeans for the portions. Tables 7 and 8 show the calculation, first of a slope based on the residual-selected midmeans of the portions, then of an overall level based upon the midmean of the results of adjusting each value for the fitted slope.

In this instance, repeating the process with the new residuals based on the fitted tilt does not produce anything new, since -- as table 9 shows, when compared with table 7 -- the order of the residuals within portions was not changed (this time) by the tilt fitting. (It remained 1921 < 1918 < 1919 < 1920 and 1929 < 1928 < 1927 < 1926.)

The final result is, then

$$3{,}804 + 74.4(x-1923.5)$$

in place of

$$3{,}851 + 63.2(x-1923.5)$$

table 7 of chapter 7

Calculation of a Tilt Based Upon Residual-chosen Midmeans of Portions (residuals from y-means)

(1) Residuals, ordered within each portion.

(2) Corresponding values of x = date.

(3) Corresponding values of y = imports in millions of dollars.

(4) and (5) Observations to be used for portion midmeans.

(6) and (7) Portion midmeans.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
-1184	1921	2,509				
-472	1918	3,031	1918	3,031	1918.5	3,468
+337	1919	3,904	1919	3,904		
+1648	1920	5,278				
-160	1929	4,039				
-44	1928	4,091	1928	4,091	1927.5	4,138
113	1927	4,185	1927	4,185		
422	1926	4,431				

(8) $\text{Tilt} = \dfrac{4,138 - 3,468}{1927.5 - 1918.5} = 74.4.$

table 8 of chapter 7

Calculation of a Typical Value Based on the Midmean of Tilt-adjusted Values

(1) x = year.

(2) y = imports in millions of dollars.

(3) b(x-1923.5) = adjustment for fitted tilt; b = 74.4.

(4) Column (2) minus column (3) = imports adjusted for tilt.

(5) Values of column (4) ordered.

(6) Calculation of midmean of these values; x's mark wing values.

(1)	(2)	(3)	(4)		(5)	(6)
1918	3,031	-409	3,440	.	2,695	x
1919	3,904	-335	4,239	:	3,225	x
1920	5,278	-260	5,538	:	3,440	x
				:		
1921	2,509	-186	2,695	:	3,572	
1922	3,113	-112	3,225	:	3,630	
1923	3,792	-37	3,829	:	3,756	$\frac{22,827}{6} = 3,804$
1924	3,609	+37	3,572	:	3,829	
1925	4,227	+112	4,115	:	3,925	
1926	4,431	+186	4,245	:	4,115	
				:		
1927	4,185	+260	3,925	.	4,239	x
1928	4,091	+335	3,756	:	4,245	x
1929	4,039	+409	3,630	:	5,538	x

table 9 of chapter 7

Calculation of a Tilt Based Upon Residual-chosen Midmeans of Portions (residuals from fit in Table 8, i.e., values in column (4) of that table less 3,804)

(1) Residuals, ordered within each x-defined portion according to their values.

(2) Corresponding values of x = date.

(3) Corresponding values of y = imports in millions of dollars.

(4) and (5) Observations to be used for portion midmeans.

(6) and (7) Portion midmeans.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
-1109	1921	2,509				
-364	1918	3,031	1918	3,031		
435	1919	3,904	1919	3,904	1918.5	3,468
1734	1920	5,278				
-174	1929	4,039				
-48	1928	4,091	1928	4,091		
121	1927	4,185	1927	4,185	1927.5	4,138
441	1926	4,431				

(8) $\text{Tilt} = \dfrac{4,138 - 3,468}{1927.5 - 1918.5} = 74.4.$

In this example, the change is small, only 109 million dollars for 1918, where the change is largest.

There are, however, many instances where one ought to desire to avoid the possible dominance of a single very unusual value -- or a few such values -- in determining the fitted line. In many of these instances, going through these further steps -- either once or more than once -- can gain us much.

7D. Introduction to bends and unbending

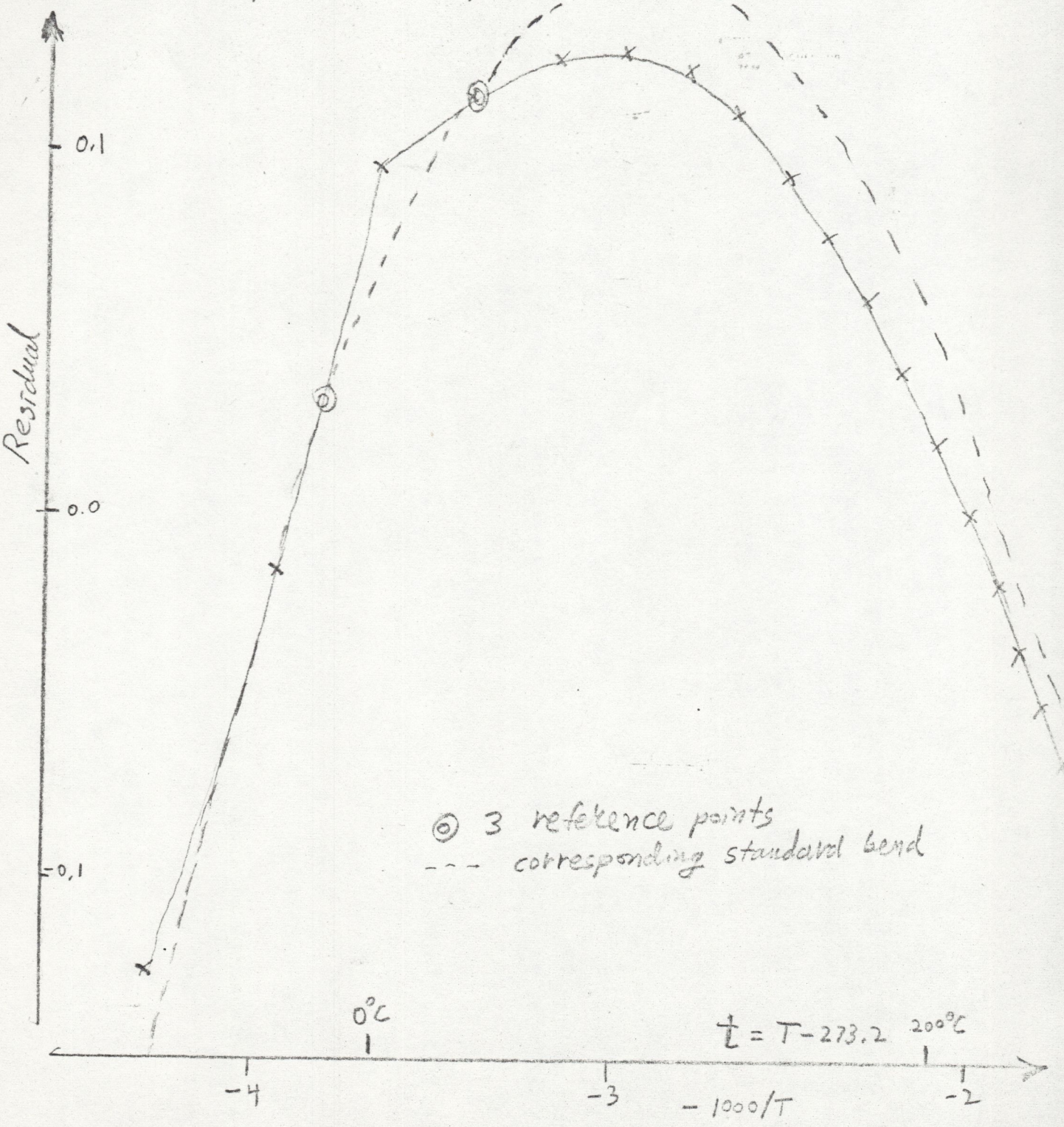
Now that we have learned to deal with tilt, it is time to go further. Untilting things does not necessarily make them flat; we may need to unbend them also. (And even simple unbending may not be enough.) The easiest step to take in this direction is to choose some simple standard sort of bending and plan to remove it by subtraction, just as we chose a simple sort of tilting -- along a straight line -- and removed it by subtraction. Just as we cannot expect the result of untilting to be unbent, we cannot expect the result of unbending to be untwisted. We can hope that unbending will help us to see better what goes on beyond or below the phenomena we are taking account of.

Plot **10** repeats plot 30 of chapter 2, showing an untilted form of the log vapor pressure of water, with the addition of:

- three selected points
- a standard bend through these points.

plot 10 of chapter 7

Residuals of log vapor pressure of water from a straight line. (Repeat of plot 30 of chapter 2.)



Plot **11** shows the further residuals obtained by subtracting this standard bend from the points. We now see considerably more than we did before. The break at 0°C, corresponding to the difference between liquid water and solid water is much more prominent, and we see that our bend, chosen to fit more or less overall, does not match the bend for liquid water very closely.

To see more, we need to fit a bend to the liquid water portion of the data. It would be nice if we could do this by merely fitting another standard bend on top of the one we have already fitted. Gaining the ability to do this would be a good reason for choosing a particular family of standard bends.

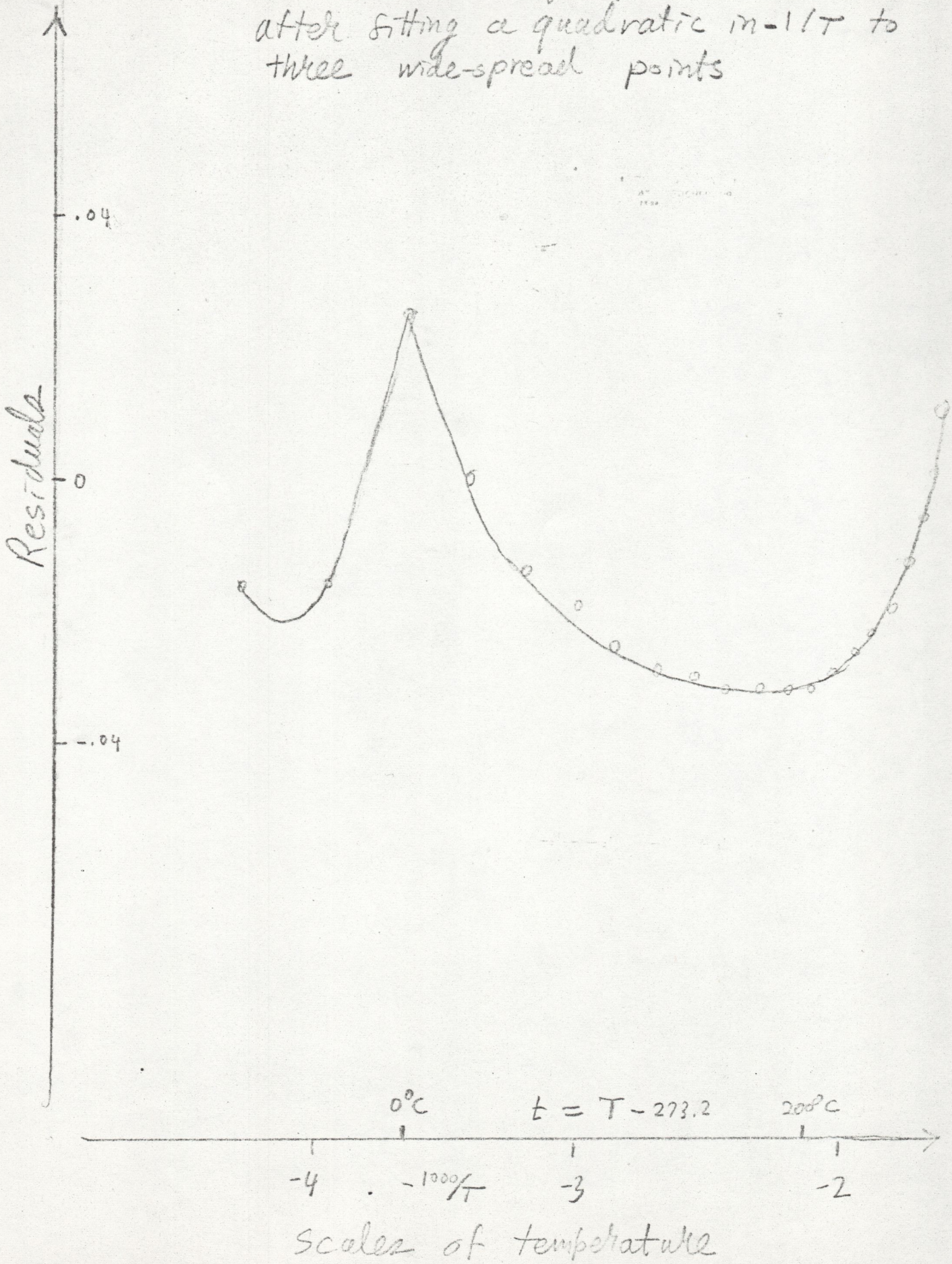
In fact, the standard bends we are using do have this property. Choosing three points on the liquid water curve and subtracting the standard bend through them gives plot **12**. Here we see a relatively straight portion of the data from 0°C up to about 300°C with a rather sharp deviation at each end.

If we select a standard bend to fit the solid water portion of the curve -- doing this is left to the interested reader -- we find that it fits quite well, too. Thus we stand at the following position:

- a combination of standard tilt and standard bend fits the solid water data quite well.
- a different combination of standard tilt and standard bend fits the liquid water data quite well from 0°C to 250°C.

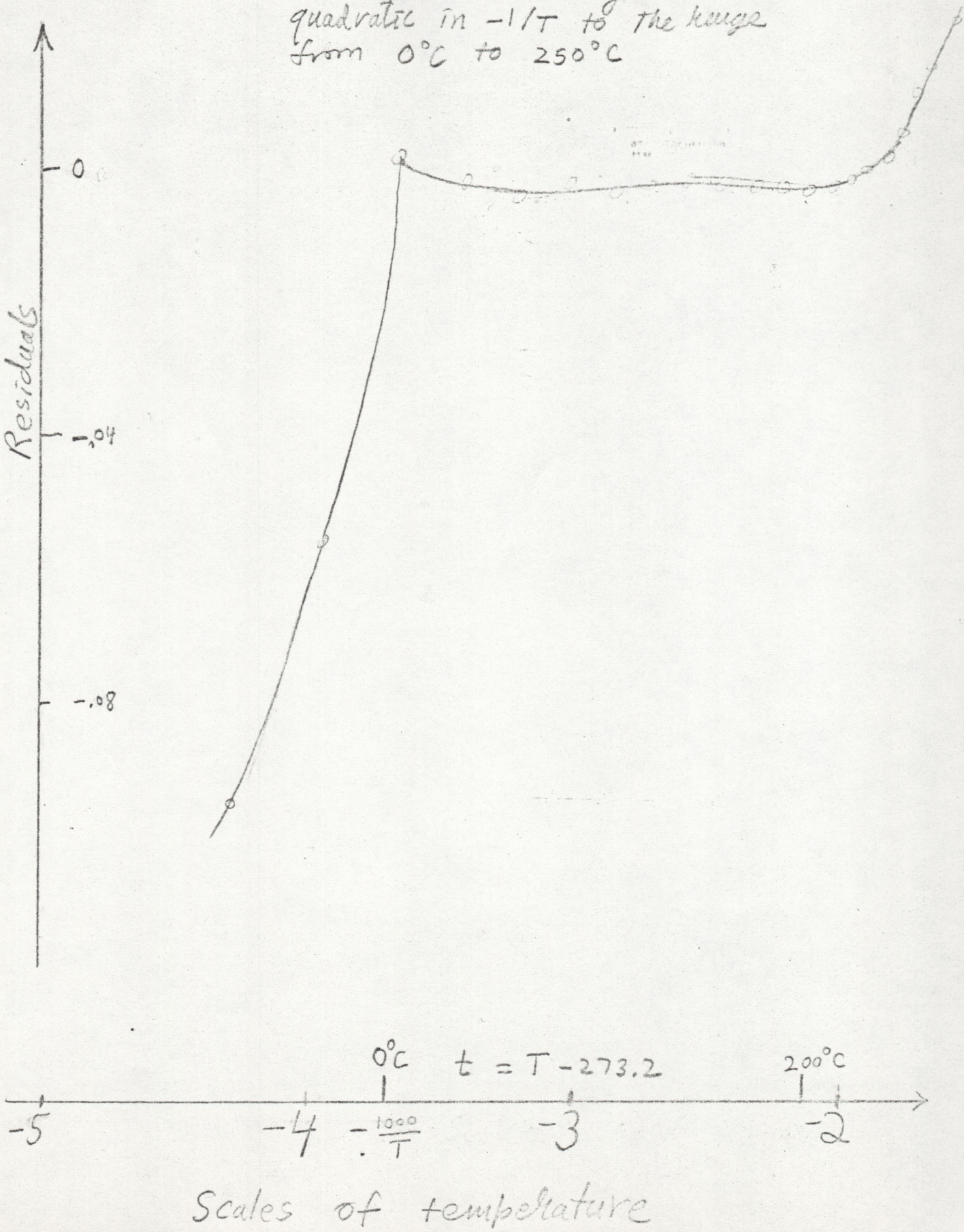
plot **11** of chapter 7

Residuals of log vapor pressure of water after fitting a quadratic in $-1/T$ to three wide-spread points



plot 12 of chapter 7

Residuals of log vapor pressure of water after fitting a quadratic in $-1/T$ to the kurge from 0°C to 250°C



- something begins to happen to water above 250°C.

What is it?

So much of introduction and introductory example. Before we turn to the question of how to fit bends, we need only explain what family of bends we chose as standard bends. Any such family needs to be carefully related to the family of standard tilts that will be used with it. We must be prepared to face U-shaped relations of y to x whose low points fall at different values of x. Yet we want to have very simple standard bends, bends that can be easily described by single numbers. The only way to meet this aim is to use standard tilts to move the low point of the bend back and forth. This means that we depend on sums and differences of standard bends and standard tilts to provide U-shaped partial descriptions with low points wherever we need them.

It would be nice for the subfamily of U-shaped partial descriptions with a given low point to offer the same choices as that for any other low point. This happens for only one choice of a simple family of standard bends, the same choice, fortunately, which most of us would find simplest:

- quadratic expressions in x.

What happens when we combine a specific quadratic and a specific straight line? A simple example is

$$\begin{aligned}[(x-10)^2 + 25] + [6x-19] &= x^2 - 20x + 125 + 6x - 19 \\ &= x^2 - 14x + 106 \\ &= [(x-7)^2 + 57]\end{aligned}$$

which shows that, in this case, we have moved the low point both sideways -- from $x = 10$ to $x = 7$ -- and upward -- from $y = 25$ to $y = 57$. Except for this rigid displacement, the shape has remained the same. In general, because

$$[(x-a)^2 + c] + [2(a-b)x + a^2-b^2-c+d] = [(x-b)^2 + d]$$

what happens is exactly similar, a rigid displacement, moving the low point to $x = b$, $y = d$. Clearly, we can choose first d and then c to put the low point where we wish.

7E. Assessing bend free of tilt

Let us turn, then, to the fitting of a standard bend. The easiest case to treat, of course, is one where three points are given. Suppose, then, that we are given these three points:

x	y
3	7
5	4
8	11

How are we to assess the amount of bend corresponding to these points? We surely wish that our assessment of bend should avoid any response to pure tilt. What is an easy way to do this?

One easy approach is to use the three points to calculate two different tilts, and then depend on the difference between these tilts to assess the bend. For our example, we have -- for data and for x^2, a standard bend:

x	y	tilt	x	x^2	tilt
3	7		3	9	
		-1.50			+8.00
5	4		5	25	
		+1.33			+13.00
8	8		8	64	
Diff:		+2.83			5.00

Since the observed difference in tilt is

$$\frac{2.83}{5.00} = .567$$

as large for the data as for x^2, the natural assessment of bend corresponds to a bend described by $.567x^2$. Applying this we find:

x	y	$.567x^2$	$y-.567x^2$	tilt
3	7	5.10	1.90	
				-6.04
5	4	14.17	-10.17	
				-6.03
8	8	36.27	-28.27	
	(data)	(bend)	(residual)	

Clearly this process has provided three residuals related by pure tilt.

In doing this example, we used x^2 as an example of a standard bend. There was no need for making this specific choice. If, as another example, we had chosen $(x-5)^2$, we would have found:

x	y	tilt	x	$(x-5)^2$	tilt
3	7		3	4	
		-1.50			-2.0
5	4		5	0	
		+1.33			+3.0
8	8		8	9	
Diff:		2.83			5.0

again finding 0.567 as the coefficient, here of $(x-5)^2$. Continuing, we would find

x	y	$0.567(x-5)^2$	$y - 0.567(x-5)^2$	tilt
3	7	2.27	4.73	
				-0.37
5	4	0	4.00	
				-0.37
8	8	5.10	2.90	

again finding three residuals which represent pure tilt. (We expect this pure tilt to be different from that found by subtracting $0.567x^2$.)

Clearly either approach leads to a satisfactory answer and is equally useful. Equally, we could have used $(x-7)^2$, or $(x-\pi)^2$, etc. The only reason for one choice over another is simplicity of arithmetic. From this point of view, a choice of $(x-c)^2$ with c somewhere interspersed among the values of x leads to smaller numbers and thus, usually, simpler arithmetic. In our example, the choice $c = 5$ obviously works well.

Sometimes -- as we will soon see when c is a half-integer, but all the x's are integers -- it may simplify the arithmetic a little to use $(x-c)^2 - d$ rather than $(x-c)^2$. We are clearly free to do this, since doing this will affect the calculated slopes not at all, and will lead to the same amount of bend.

7F. A simple bend assessor

If we have more than three points, as we usually do, we need to begin with some form of summarization of the data. Again we can follow the broad general approach we used for tilt.

This time we break our observations up into five portions according to their x-values, and calculate the

apparent direction and amount of bend from the behavior of three of these: the lowest (L), the center (C) and the highest (H). The arithmetically simplest approach is to calculate the three pairs of means, which we naturally label x_L, y_L and x_C, y_C and x_H, y_H respectively. Then we ask these three points what bend they indicate.

As before, it remains only to choose the sizes of the three portions that we are to use. Again a rule of thumb serves us well. This time each recommended portion, if possible:

- starts from an end -- L or H -- or -- C -- is centered around the median.
- extends at least 1/5 of the way from min x to max x.
- includes at least 1/7 of the observations.
- is otherwise as short as possible.

Table 13 illustrates the computation, using the means of the three selected portions, for the 1918-1929 data on annual imports. In this instance the extreme value for 1920 appears to be a substantial contribution responsible for the large effect, and we are impelled to bring in midmeans.

Before we do this, however, we should do a simple check on the effectiveness of our bend assessment. We found that the bend appeared to be described by

$$20.5[(x-1923.5)^2 - 0.25]$$

To check up, we should examine

$$y - 20.5[(x-1923.5)^2 - 0.25]$$

table 13 of chapter 7

Evaluation of Apparent Bend in the Value of U. S. Imports

(1) x = date.

(2) y = value of imports in millions of dollars.

(3) values of $(x-1923.5)^2 - 0.25$.

(4),(5), and (6) means of columns (1), (2) and (3) for the low, center, and high portions.

(7) tilts of column (5) against column (4): e.g.

$$\frac{3,685-4,071}{1923.5-1919} = -85.8.$$

(8) tilts of column (6) against column (4): e.g.

$$\frac{-11-(8\frac{2}{3})}{1923.5-1919} = -4.37.$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1918	3,031	30					
1919	3,904	20	1919	4,071	$20\frac{2}{3}$		
1920	5,278	12					
1921	2,509	6				-85.8	-4.37
1922	3,113	2					
1923	3,792	0	1923.5	3,685	1		
1924	3,609	0					
1925	4,227	2					
1926	4,431	6				+93.3	+4.37
1927	4,185	12					
1928	4,091	20	1928	4,105	$20\frac{2}{3}$		
1929	4,039	30					

(9) ratio of differences of tilts:

$$c = \frac{(+93.3) - (-85.8)}{(+4.37) - (-4.37)} = 20.5 = \text{measure of bend}$$

and see whether, by the same criterion, it seems free of bend. Table 14 shows the calculation. We see that the two slopes in column (7) are the same -- to the precision of our calculations -- indicating zero bend in column (1), at least when examined in this way. Thus our computations check.

If we take y-selected (easy) midmeans instead of means, we find:

(4)	(5)	(6)	(7)	(8)
1919	3,904	20		
			-45.3	-4.44
1923.5	3,700	0		
			+86.9	+4.44
1928	4,091	20		

for which

$$c = \frac{(+86.9) - (-45.3)}{(+4.44) - (-4.44)} = 14.9$$

is appreciably lower than before.

To do a possibly better job, we could subtract

$$14.9[(x-1923.5)^2 - 25]$$

from each observed value, and then proceed to fit a tilt using midmeans. After subtracting this tilt, we could then use the resulting values to select new midmeans of the three "fifths", thus starting over again. In the present example we could not expect a large change to result from doing this and finding new bends, new tilts, and a new constant. In other situations carrying out such a chain of steps could be important.

table 14 of chapter 7

Values of U. S. Imports After Bend Removal By Means

(1) x = date.

(2) $y - 20.5[(x-1923.5)^2 - 0.25]$, where y = value of imports.

(3) values of $(x-1923.5)^2 - 0.25$.

(4),(5), and (6) means of columns (1), (2), and (3) for the low, center, and high portions.

(7) tilt of column (5) against column (4).

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1918	2,416	30				
1919	3,494	20	1919	3,647	($20\frac{2}{3}$)	
1920	5,032	12				
1921	2,386	6				+4.0
1922	3,072	2				
1923	3,792	0	1923.5	3,665	(1)	
1924	3,609	0				
1925	4,186	2				
1926	4,308	6				+3.6
1927	3,939	12				
1928	3,681	20	1928	3,681	($20\frac{2}{3}$)	
1929	3,424	30				

(8) Difference in tilts: 4.0 - (3.6) = 0.4, close to zero.

7G. Lessons about fitting in general

Here in chapter 7 we are again being a little serious about fitting. As a result we have come upon examples of some very general characteristics of all known fitting processes - characteristics that could only be avoided by the use of extrasensory perception or some similar miracle.

When we come to fit, whether our purpose is to assess or to clear out of the way, we must (unless we are operating graphically in a freehand way) choose a detailed fitting procedure. We would like to choose this procedure so as to listen only to the sort of message we are seeking. This we can never do. The best that we do is to make our procedure deaf to certain specific sorts - a few, not many - of messages that we do not wish to hear.

Consider a man who is concerned with the slope of y as a function of x - a man who, as "slope" suggests, is thinking of y depending on x in a straight-line way. What can he do to listen only to straight-line messages? There is at least one thing he can do - one thing we all do, usually without conscious thought. He can avoid completely any effect of the general size - the general value - of y.

In 3A we took

$$\frac{y_H - y_L}{x_H - x_L}$$

as our measure of slope. If each and every individual y were increased by, say, 365, both y_H and y_L would be increased by

365 and, since $y_H - y_L$ would not change, the fitted tilt would not change at all. To the message "all y's are increased (or decreased) by the same amount" our slope assessment is completely deaf. This is good; this can be done; not to do it would usually be close to shocking.

But what if we change the y's in a way that is not represented by a straight line? Suppose we change the y's in proportion to $(x-\min x)^2$ where min x is the smallest x represented in the data? This change is far from a straight-line change. A purist looking for straight-line slopes might wish not to see it. But, since y's for high x's are certainly increased more than y's for low x's, the tilt we assess using this approach -- or any reasonable approach -- will surely appear positive.

There was nothing particular about x^2, we could have said the same thing about any steadily (or jerkily) increasing function of x. An assessor of slope cannot avoid being sensitive to very many sorts of changes. Only a few special changes will leave its result unaffected. (This is one of the reasons why we speak of a <u>tilt</u> assessor rather than of a <u>slope</u> assessor.)

One of the arts of careful fitting is arranging that the few changes that leave the result unaffected include those whose possible influence on the result we most wish to avoid.

The assessment of bend offers another example. There we took the following steps:

- first find the tilts of y against x and of $(x-c)^2 + d$ against x for two different intervals,

- then find the differences of these tilts and form their ratio.

The whole shape of this calculation is designed to reject what we most want to reject: any straight-line (= wholly unbent) behavior. A constant contribution to all y's disappears at the first step, where we take the tilts. A steady rate of increase or decrease of y with x sneaks through the first step -- raising or lowering the two tilts of y against x by exactly equal amounts -- but is caught, and rejected, at the second step -- where the difference of these two tilts is taken.

Almost any other dependence of y on x will, however, provoke some response in the result from this bend assessor. Care allows us to eliminate responses to a few sorts of change. Nature insists upon responses to almost all others.

The same relation of the attainable to the desirable was already present in 1D. Once we both admit that there may be "wild" values and decide that we should like to neglect them, it is easy to say what we would like to have, namely:

- a typical value that is unaffected by a change in any ONE of the observations from which it is computed.

It is equally easy to see why it is impossible to have such a utopian typical value. Consider increasing one y after another by an equal amount, say by 100 units, until all have been increased. Each step - since only one value changes - would not change our utopian typical value at all. The overall change, however, in which each and every y increases by 100 is one that must affect anything worthy of the name of typical value. There can be no such utopian typical value.

If we cannot have all we would like, how much can we have? Both median and midmean have a large part of what we want, namely:

- if we change just one number more and more, the time comes when further change (of this one value in this one direction) has no effect on our typical value.

So much we can have; all we might dream of we cannot.

There is a third lesson to be drawn from our experiences. When we are to fit things of ascending complexity, such as a constant and a tilt - or a constant, a tilt and a bend - we are likely to find it well to begin by fitting the most complex thing and to fit the less complex things afterwards. This is particularly true when we are realistic enough to take specific precautions to deal with erratic, or individualistic values. Thus, we found it natural to fit the bend before fitting the tilt, and to fit the tilt before fitting the constant, while in 7C we fitted the (midmean) tilt before fitting the midmean itself.

7H. How far have we come?

The emphasis of this chapter has been on technique, on how we can tame arithmetic to do almost as well as a graphic approach. The great gains in doing this are:

- standardization, so that it is easy to say what was done,

- automation, so that the computer can carry the analysis further before we have to look at it,

- generalization, so that we can do more complicated things than it is easy to do graphically.

Such arithmetization of what we learned to do graphically is not done with the intention of elimination of graphs. Rather it is done to let arithmetic, especially on an automatic computer, take us further down the road. As a result, we can more frequently get all our graphs at one time, rather than one after another. The fewer "passes" at the computer that are needed to reach a given penetration, the better off we are.

More specifically we have looked at the assessment of tilt and bend by arithmetic techniques. The following points prove important:

- the comparison of typical values for lower and upper "thirds" is an effective way to assess tilt.
- it is both possible and useful to use midmeans in this process, though we must be careful as to which midmeans we use - indeed, we must be prepared for repeated (iterative) calculation, at least in some instances.
- looking for bend is also not difficult: we can use the means or midmeans of three spaced "fifths" of the observations.
- a tilt need not mean a slope: almost any steady tendency for y to increase as x increases will be found by a tilt assessor and given a number as if it were a slope. (This will happen to some dependences in which the tendency is far from steady, too.)
- fitting is always assessing the numbers in a partial description of fixed form; the general case is like tilt and slope - a fit will provide a more or less close

partial description, but no certainty that the form of this description is at all "correct".

- the most we can do in fitting is to avoid a very few forms of interference: We can fit a tilt in a way not affected by a constant; we can fit a bend in a way not affected by a constant or a tilt. We can go little further.

- in practice, we often fit the more complex or recondite aspects first -- first, not last; thus we fit bend before tilt before level.

* * * * *

7K. <u>How to add</u>

Addition, like other operations requiring care, is more or less painful to all of us. Anything we can do to reduce stress and strain, or to increase accuracy, cannot help but be worthwhile. Strangely enough there is a simple thing to do that helps to do both these things when adding columns of figures - something called "bookkeeper's addition" by some - something far too rarely mentioned.

All - or nearly all - of us have been taught to "carry" by either remembering or writing down in an unobtrusive or scratchy way whatever needs to go over from one column to the next. As a result, carrying is usually the most painful - and the least accurate - part of our addition. This need not be.

In "bookkeeper's addition" we proceed as follows:

- add up each single column of figures separately, writing down the total, appropriately, at the bottom of the column.

- add up the totals thus obtained to find the answer to the whole sum.

While we could put each column total on a separate line thus

```
    59
   45
  3,7
42,
46,209
```

(the column totals - 59,45,37,42 - are for (2) of table 13) it is

usually easier to let the column totals take up only two or three lines as in

```
   3759
42 45
------
46 209
```

This means reusing a line as soon as there is space to do so.

In any event, checking the addition can be easily done single column by single column.

For the writer, the great advantage of bookkeeper's addition comes when he is tired -- and consists mainly of lowered strain.

*****

7P. IEDA processor features

7R. A few words about measurement

Section 7A starts with the assumption that the way to assess the tilt of y against x is to ask: "For what b is y - bx apparently untilted?". This flat statement appears more natural to some readers, and less natural to others. The present section provides some background about -- and, hopefully, some insight into -- what is going on in measurement in general.

It is all too easy to come to believe that measurement is a matter of great depth, or esoteric, or even sacred. Each of these views has misled many people. We would like to avoid as many of them as we can.

Enrico Fermi, the Nobel Laureate physicist, is reliably reported to have once said: "Measurement is just the making of fine distinctions".

Norman Robert Campbell, who thought as hard and well as anyone about the measurements of classical physics wrote: "Now the devising of methods for judging > is the chief problem of the experimental art. For the range of possible methods is enormous: any effect, however remote and indirect, of change in a magnitude provides a possible method; and the most ingenious experimenter is he who can see ways of using very remote effects". (An account of the principles of measurement and calculation: pp. 41-42.)

The one essential of all carefully axiomatized schemes of measurement (see page 1 of Vol. 1 of J. Math. Psychol. for

an unCampbellian instance) is the notion of >, of an ability to compare -- and thus to recognize order.

As Campbell puts it on page 8 of the same book: "The conception of a magnitude is inseparable from that of the order characteristic of it. It is natural, therefore, to regard as the same magnitudes, or as magnitudes of the same kind, properties that invariably have the same order".

One basic thing in establishing how a specific property -- for example, tilt -- is to be measured, is to decide how things are best compared. In the case of tilt, the things we are computing can be different enough in kind to make it worthwhile to recognize this difference in setting up our scheme of measurement.

The tilt of one straight line is easily compared with that of another straight line. The same is not true of two wiggly curves or two scattered trails of points -- to say nothing of two point clouds that are only somewhat elongated. Moreover, we know how to add tilts of lines. Since + and > together are enough, when well behaved (as they are here), to determine solid assignments of numbers, we can -- and should -- start with a quantitative notion of tilt for straight lines, and then extend this to more general patterns.

Our problem then -- given a wiggly curve, a scattered point trail, or a moderately elongated point cloud -- is to compare the tilt of this pattern with the tilt of various straight lines. We can do a low-to-medium grade job of this by drawing a straight line "through" the pattern before us.

This process is often useful and ought never to be neglected. However, we can do better. Since better is available, our fundamental definition should take advantage of it.

The best way to compare the tilt of a wiggle, point-trail, or point-cloud with the tilt of a straight line is just to subtract the line from the pattern being studied. The eye is best at comparing with a horizontal line -- and, it may be possible, once the line is untilted, to increase the magnification with which we look at the result. In practice or in theory we ought to define the tilt of a pattern as the slope of a line which, when subtracted, makes the pattern look untilted.

References for this section:

N. R. Campbell 1928. An account of the principles of Measurement and Calculation. 293 pp. London: Longmans, Green. (This book is harder to find than Campbell's Physics, the Elements, Vol. 1, which was reprinted by Dover as Foundations of Science: The Philosophy of Theory and Experiment, New York, 1957.)

For a recent symposium on measurement, see: Measurement: Definitions and Theories. 274 pp. New York: John Wiley & Sons, Inc., edited by C. West Churchman and Philburn Ratoosh.

Recent attention to what is called simultaneous conjoint measurement began with a rather technical paper, whose opening sections may prove readable, namely:

R. Duncan Luce and John W. Tukey 1964. Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement, J. Math. Psychol., 1-27.

7S. <u>When computing is no pain</u>

The process of dividing the data into thirds, typifying the end thirds, and using the two typical points to fix a tilt:

- is rather crude, but effective.
- presumably can be replaced by more polished processes requiring more effort.
- is not likely to be so replaced for hand calculation.

In this section we look at one more polished technique. Its basic idea is simple. Any two points -- (x_1,y_1) and (x_2,y_2) -- define a tilt

$$\frac{y_2-y_1}{x_2-x_1}$$

which can reasonably be given weight

$$(x_2-x_1)^2$$

(Under suitable conditions and by techniques outside our scope, one can show this is the natural weight. It is enough here to say that it is reasonable appearing.) To combine these weighted tilts, we naturally use a midmean.

((There is some interest in looking at the weighted mean, which takes the form

$$\frac{\frac{(y_1-y_2)}{(x_1-x_2)}(x_1-x_2)^2 + \frac{(y_1-y_3)}{(x_1-x_3)}(x_1-x_3)^2 + \ldots}{(x_1-x_2)^2 + (x_1-x_3)^2 + \ldots}$$

$$= \frac{(y_1-y_2)(x_1-x_2) + (y_1-y_3)(x_1-x_3) + \ldots}{(x_1-x_2)^2 + (x_1-x_3)^2}$$

$$= \frac{(y_1-\bar{y})(x_1-\bar{x}) + (y_2-\bar{y})(x_2-\bar{x}) + \ldots}{(x_1-x)^2 + (x_2-\bar{x})^2 + \ldots}$$

(where the last step may not be obvious) a form that has been used rather frequently.))

Table 15 shows the process in action. We find that the slopes thus found fit in well with our earlier results, for we have the following tilts:

means of thirds:	63.2
midmeans of thirds:	74.4
weighted skeleton:	72
weighted midmean:	70

Since the number of pairs of observations rises rapidly as the number of observations increases, we would like to save calculation. It is natural to do this by selecting a few points and using only the tilts found by connecting each of these few points to each point of all. Unfortunately this does not seem to work as well as one would hope. Table 16 shows this for a few choices and the familiar example. If we are to short cut, the methods of tables 3 and 7 still look good, especially the latter.

table 15 of chapter 7

The Weighted-Tilt Process Applied to the Data of table 3

a) THE TILTS, $(y_j-y_i)/(x_j-x_i)$ IN TRIANGULAR FORM

	873	1124	-174	21	153	-70	171	175	128	106	91
	1374	-698	-264	-28	-258	-54	75	35	21	14	
	-2769	-1082	-495	-417	-210	-141	-156	-148	-138		
	604	642	367	430	384	279	226	191			
	679	248	371	330	214	163	132				
	-183	218	213	98	60	41					
	618	411	192	120	86						
	204	-21	-45	-47							
	-246	-170	-131								
	-94	-73									
	-52										
(Wts)	(1)	(1)	(4)	(9)	(16)	(25)	(36)	(49)	(64)	(100)	(121)

b) ORDERED TILTS (WITH WEIGHTS)

1374(1)	367(9)	175(64)	75(49)	-52(1)	-174(9)
1124(4)	330(16)	171(49)	60(25)	-54(36)	-183(1)
873(1)	279(36)	163(36)	41(36)	-70(36)	-210(25)
679(1)	248(4)	153(25)	35(64)	-73(4)	-246(1)
642(4)	226(49)	132(49)	21(16)	-94(1)	-258(25)
618(1)	218(4)	128(81)	21(81)	-131(9)	-264(9)
604(1)	214(25)	120(16)	14(100)	-138(81)	-417(16)
430(16)	213(9)	106(100)	-21(4)	-141(36)	-495(9)
411(4)	204(1)	98(16)	-28(16)	-148(64)	-698(4)
384(25)	192(9)	11(121)	-45(9)	-156(49)	-1082(4)
371(9)	191(64)	86(25)	-47(16)	-170(4)	-2769(1)
(Sum) (67)	(226)	(582)	(416)	(321)	(104)

Sum of weights 67 + 226 + 582 + 416 + 321 + 104 = 1716

Now 1716 = 429 + 858 + 429 so that hinges are at 429.5 and 1287.5, median at 858.5.

c) HINGES, MEDIAN, AND SKELETON MIDMEAN

Value	Ranks	(Inverse ranks)	
+163	1275 to 1310	(407 to 442)	U hinge
+86	842 to 866	(851 to 875)	median
-47	426 to 437	(1234 to 1245)	L hinge

$$\text{skeleton midmean} = \frac{163+86+86-47}{4} = 72$$

table 15 (Cont'd) - 2

d) MIDMEAN

$$(13)(163) + (25)(153) + (49)(132) + \ldots + (16)(-28) + (9)(-45) + (12)(-47) = 60{,}233$$

$$(13) + (25) + (49) + \ldots + (9) + (12) = 858$$

$$\frac{60{,}233}{858} = 70$$

e) EXAMPLES (Wts)

$$\frac{3{,}904-3{,}031}{1919-1918} = 873 \qquad (1)$$

$$\frac{5{,}278-3{,}031}{1920-1918} = 1124 \qquad (4)$$

$$\frac{2{,}509-5{,}278}{1921-1920} = -2769 \qquad (1)$$

$$\frac{2{,}509-3{,}031}{1921-1918} = -174 \qquad (9)$$

table 16 of chapter 7

Results of Using a Few Selected Points to Combine With All to Give Tilts Which Then Give a Weighted Midmean

Selected points	Tilt: Skeleton midmean	Tilt: Easy midmean
the mean of all 12	93	88
the midmean of 12	~100	~91
3 midmeans, one of each third	79	81
4 easy midmeans (= medians), one of each quarter	54	52

Other approaches	tilt
table 3	63
table 7	74
table 15, skeleton	72
table 15, midmean	70

*****

7Y. Problems

Problems for hand arithmetic

1. Consider the colorimetric data in problem 3 of chapter 2. Fit a tilt arithmetically. Plot the residuals. Discuss.

2. Consider the carbon content of 36 clays as given in table 39 of chapter 2. Which clays fall in each of the three portions according to the simple rules of section 7A? Fit tilts to these values, using first means and then midmeans. Find and plot residuals.

3. Assume at least three values for the amount of slowly decaying radioactivity in the two left-hand columns of table 40 of chapter 2. (For which manner of expressing amount of radioactivity?) Calculate the corresponding tilts using midmeans. Plot the residuals. Discuss.

4. Assess the bend in the logarithms of the vapor pressure of mercury for temperatures from 0°C to 180°C, using the data given in table 41 of chapter 2. After subtracting the bend, assess the tilt and find the residuals from the partial description combining bend and tilt. Plot these residuals. Discuss. Apply the same partial description to the full range given, finding and plotting residuals. Discuss.

5. Take the 1952 election results in table 42 of chapter 2 as x and the 1956 results as y. Fit a tilt arithmetically. Find the residuals and plot them. Discuss the result.

Extra-hard problems

31. Show that a family of standard bends which, when combined with standard straight-line tilts, offers essentially equivalent choices for any fixed low point must be describable by quadratic functions of x.